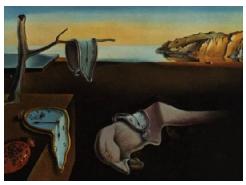
View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Papers in Economics



Journal of Statistical Software
November 2008, Volume 28, Book Review 2. http://www.jstatsoft.org/

Reviewer: Markus Schmidberger
Ludwig-Maximilians-Universität München

Introduction to Machine Learning and Bioinformatics

Sushmita Mitra, Sujay Datta, Theodore Perkins, and George Michailidis
Chapman & Hall/CRC, Boca Raton, Florida, 2008.
ISBN 978-1-58488-682-2. 384 pp. USD 79.95.

Introduction

Machine learning (Hastie *et al.* 2001) is a sub-set of artificial intelligence and deals with techniques to allow computers to learn. Bioinformatics is the application of information technology to the area of molecular biology. For both disciplines there are already many books introducing the basic concepts and there are some books presenting machine learning applications in the area of bioinformatics. The book under review covers essentially topics of machine learning and bioinformatics and demonstrates the usefulness of statistical methods in well-documented bioinformatic examples. In the first part, the book teaches basic concepts of machine learning and introduces essential biological aspects. In the second part, the authors close the gap between the two disciplines and present a lot of interesting examples. It is a well-structured book that is a good starting point for machine learning in bioinformatics.

The book has 12 chapters which can be grouped in two parts: Chapters 1–6 cover the introduction and the basics. Chapter 7 describes the connection between machine learning and bioinformatics. Chapters 8–12 include well-selected and ongoing bioinformatic examples and are primarily written by co-authors.

The authors put a lot of their teaching experiences into this book. The statistical basics are illustrated with well-chosen and popular examples. Every chapter (except Chapter 2) ends with exercises and references.

Book contents

Chapter 1 starts with the historical development of the amount of biological data and introduces connections and dependencies between statistics, biology, and computer science.

Chapter 2 covers a brief review of the essential concepts in cellular and molecular biology to motivate the biological problems. It starts with the basic unit of living organisms (cells),

DNA and biological regulation systems and ends with measurement technologies for molecular biology. This chapter requires no biological knowledge but is important for the following chapters. Biologists can skip this chapter.

Chapter 3 introduces the statistical basics: probability theory, stochastic processes, hidden Markov models, etc. This chapter requires no statistical knowledge but is important for the following chapters. Statisticians can skip this chapter.

Classification, also known as supervised learning, is required in a lot of different disciplines. Through the power of computers new classifiers were developed in the last years: classification trees, support vector machines, boosting, etc. Chapter 4 describes the general statistical decision framework for classification and discusses a number of classification methods.

In classification it is often reasonable to use dimension reduction methods to simplify the data. Chapter 5 covers components analysis and multidimensional scaling for dimension reduction. Furthermore, a problem formulation and established algorithms for cluster analysis are described. Cluster analysis is the classification of a set of objects in different groups and a form of unsupervised learning.

In Chapter 6 the authors introduce computational intelligence in a very brief way. The chapter ends with a long discussion about opportunities and applications for computational intelligence in different areas of bioinformatics.

In Chapter 7 the authors describe connections between bioinformatics and machine learning. For this they select three well-known problem areas: sequence analysis, gene expression analysis, and inference of expression data.

Machine learning and computer vision algorithms can be used for automated electron-density map interpretation. Chapter 8 gives a short introduction into structural analysis and the automatic interpretation of 3D protein images. This chapter is more algorithmically based and describes four interpretation methods in detail.

Based on the high-dimensionality of gene expression data Chapter 9 deals with biclustering as a data mining strategy. Biclustering allows for clustering of the rows (genes) and columns (samples), simultaneously. The state-of-the-art algorithms are presented, benchmarked and discussed for the gene expression *Yeast* data.

Chapter 10 focuses on the problem of classification based on microarray data. They describe an approach using all genes—rather than eliminating more than 90%—for classification: Bayesian binary classification models for prediction based on a reproducing kernel Hilbert space. The chapter concludes with a benchmark over three publicly available datasets.

iTRAQ is a technique used to identify and quantify proteins from different sources in one single experiment. Chapter 11 very briefly presents a statistical model to describe sources of variation in the data. The model allows for testing of the hypothesis of differential expression of proteins.

The book ends with Chapter 12 with a review of automated techniques for classification of database search results and the application to proteomics.

Conclusion

Actually, there is one other book which has a very similar content and structure: Baldi and Brunak (2001). Based on the clear structure, the basic chapters and the understandable

examples the book under review seems to be suitable as a textbook for a graduate course. Up to the time of review, there were no solutions for the examples available and no further information on the website.

The authors concentrated on statistical aspects and motivating examples, but they neglected the computational aspects. This means there are only a few printed code examples or details concerning software and implementation aspects. Most of the described machine learning techniques are already implemented for example in packages for the R system (R Development Core Team 2008) and adapted bioinformatic applications are available in the Bioconductor project (Gentleman *et al.* 2004).

The authors selected an interesting and very comprehensive way to present mathematical aspects without using the theorem-proof syntax. Especially the very well-known examples help to understand the theory.

In summary, in the book under review the authors introduce the reader to machine learning and bioinformatics. Using many popular examples, the statistical theory becomes comprehensible and bioinformatic examples motivate to apply the concepts to real data.

References

Baldi P, Brunak S (2001). *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge. ISBN 0-262-02506-X.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Li FLC, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004). "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, **5**, R80. URL http://genomebiology.com/2004/5/10/R80.

Hastie T, Tibshirani R, Friedman JH (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York. ISBN 0-387-95284-5.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reviewer:

Markus Schmidberger
Chair of Biometrics and Bioinformatics
Ludwig-Maximilians-Universität München
Marchioninistr. 15
D-81377 München, Germany

Currently at:
Computational Biology Program
Fred Hutchinson Cancer Research Center
Seattle, WA, United States of America
E-mail: Markus.Schmidberger@ibe.med.uni-muenchen.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 28, Book Review 2 *Published:* 2008-11-04
November 2008